

语言监测中词语构造能力的分析及其应用*

曾小兵 邱丽娜 张普 张志平 杨尔弘

北京语言大学应用语言学研究所 北京 100083

E-mail: xiaobingzeng@126.com

摘要: 2005-2009 年的国家语言资源监测工作已经积累了一些成果, 这些成果应该设法转化为语言研究、语言教学、语言信息处理的可用资源。本文将探讨语言监测中词语的内部组成及其关系, 统计并分析其中高频词语的能产性, 一方面可以为更好地析取“部件词”提供参考, 同时也可总结更多的词语结构规律。本文首先用定量统计的方法计算了高频词语的构造能力, 并分析其主要特点, 进而结合已有的类词缀研究, 分别探讨词语部件与词缀的异同, 最后从新词语提取的实践中看词缀与部件词在计算语言学中的应用。

关键词: 语言资源, 构造能力, 词汇, 结构特征, 新词获取

The analysis and application of the words' production capacity in Language Resource Monitoring

Zeng Xiaobing, Qiu Lina, Zhang Pu, Zhang Zhiping, Yang Erhong

Applied Linguistics institute in BLCU Beijing 100083

E-mail: xiaobingzeng@126.com

Abstract: The National Language Resource Monitoring and Research Center has investigated the Language situation for five years, and accumulated some resources. We attempt to use them in the field of language teaching, computational linguistics, etc. With regards to this, this article discusses the internal structures and relations of the words. It analyses the productivity of high-frequency words in statistical way, in the hope of extracting the "word component" and summing up the rules of the word's structure. Firstly, it calculated capacity of high frequency words with the method of quantification, also analyzed its main features. Then combined with the existing affixes studies, it investigated similarities and differences between "word component" and affixes. Finally, it researched the application of "word component" and affixes in computational linguistics from the practice of new word extracting.

Keywords: Language resources, production capacity, vocabulary, structure characteristics, new word extracting.

1 引言

国家语言资源监测是一项系统、持续的工程, 经过 5 年的动态监测与研究, 现阶段对于字词语的监测已经取得了一些宝贵的成果。而这些资源的合理、有效应用是当前的重要议题, 它对于深化与扩展语言资源的监测工作有重要意义, 同时也是计算语言学服务于语言学本体研究、语言教学、语言工程、辞书编纂等方面的重要体现与有益尝试。这些应用层面的探索一直在继续:

在语言生活方面, 从字词语的使用与社会生活之间的互动来看语言与社会的共变关系, 在共时监测年度语言生活状况的同时, 进一步注重反映语言生活的历时变化, 如: 从计量语言学的角度提取并收集流行语、新词语以及年度特色词等, 使语言学研究贴近社会生活。

*本文得到北京市教育委员会共建项目“汉语国际推广背景下的首都留学生教育研究”和国家语言资源监测与研究中心(平面媒体)科研项目资助, 谨致谢意! 感谢会议匿名评审的宝贵意见。

在语言教学方面,将语言资源的监测成果与汉语教学、对外汉语教学的知识动态更新相联系,探讨“以稳态词汇为核心,动态词汇模块化”的动态、数字化教学模式^[1]。

在词典编纂方面,可以将字词语做不同的等级划分,从而对字词语做定量、定序的区辨,并增加典型搭配、常用例句等属性,以满足不同规模、不同受众的词典编纂需求^[2]。

本文将探讨在语言工程中如何进一步发掘词语的作用,主要讨论词语间的内部组成及其关系,以期更好地反映字、词、语之间的关系与规律。将从高频词语的“构造能力”来看词语的生成性,并在类词缀方面做些统计分析与实践尝试,进而丰富词汇的特征描述。

2 相关研究综述

毋庸置疑,词语研究一直是语言学的重点。词语的内部结构特征很早就受到重视,赵元任^[3]就有大篇幅讲到词汇的形态类型与句法类型,朱德熙^[4]也着重谈了词的构造,许多学者如潘文国、张寿康、陈光磊等都在描述汉语构词特征上做出了贡献。在语言信息处理与计量语言学不断发展的今天,海量语料的加工技术进一步成熟,人们对于词语的计量研究更加深入。但由于语料的激增及统计手段的多样化,人们更多的是关注词语在数量上的整体分布规律或词语的变化发展状况,而较少在大规模的词语范围内去关注其内部组成及特征,这恰恰是相对而言更为重要的。

现有的一些语言资源在描述词语音、形、义等特征上做出了很大的贡献:北京大学计算语言学研究所的《现代汉语语法信息词典》充分描述了词语的语音、句法等特征;汉语的Hownet在词语语义上也有详尽的描述;台湾的中文词汇网络“在完整的知识系统下兼顾词义与词义关系的精确表达与语言科技应用”^[5]。

此外,在确定汉语常用词的范围与提取方法方面,以常用词的三个特征:全民常用性、时间的使用稳定性、构词能力强为切入点,赵小兵采用迭代算法自动识别并提取现代汉语基本词汇^[6]。黄居仁教授从分布均匀度的角度为基本词汇的验证提供了一些思考^[7]。俞士汶^[8]引入“部件词”(相当于组成词或短语的“零件”)的概念,认为:高频的“部件词”可能更接近人们通常对常用词的认知,“能逼近理想的常用词库”,期望“将基于计量研究构建的富含语法语义信息的‘部件词’库或常用词库用于汉语信息处理研究和汉语教学”。

本文主要对五年监测的高频常用词语进行了“构造能力”^{*}的考查,以三家媒体高频词交集的五年共用词表为基础,统计其在五年词语总表中的“构造能力”,这一方面是为了离析出“部件词”从而得到现代汉语中最常用的词语,另一方面也是为了反映词语之间的内部组成关系,尤其是对构词能力较强的基础词汇进行构词理据的分析,从而找到词汇发展变化的特征与规律。

3 词表资源的获取

国家语言资源监测与研究中心以流通度为标准,选取大规模的真实语料作为监测与研究的基础并建成国家语言资源监测语料库,它包括平面(报纸)、网络(新闻)、广播电视三类主流媒体,涵盖了书面语与口语(其中广播电视语料为口语转写的文本)。每年的语料量约10亿汉字次,其中平面媒体约为5亿汉字次,网络媒体约为4亿汉字次,有声媒体约为1亿字次。

^{*}构造能力其实是指构词能力,但是对于计算机自动切分的结果,总表中既有词也有短语,所以单独说是构“词”能力不确切,因此称之为“构造能力”,目的也是为了反映高频词生成其他词语的能力。

通过对语料进行自动分词与词性标注(使用中科院自动化的分词软件)后,运用统计分析的手段监测年度语言生活的字词语使用状况,并发布了2005年-2009年的《报纸、广播电视、网络(新闻)用字用语调查》。表1是从历时角度得到的五年词语的统计结果。本文基于其中的两个词表:一是三家媒体共用总词语的五年交集词表(以下简称总词表),共72641个;二是三家媒体高频词交集的五年共用词表(以下简称高频词表),共5696个,详见表1。

从监测的结果来看,总词表较好地概括了历时五年的三大主流媒体共同使用的词语的总体情况,较好地概括了大众传媒的通用、稳定词语。而高频词表则是考虑了时间(连续五年)、领域(不同媒体)、使用频率(将覆盖率达到90%的词语定义为高频词)等因素而得到的词表,可以粗略地看作是常用词语*。我们旨在考查高频词表中词语的“构造能力”,即以高频词表中的词语为基础“构件”,看其在总词表中所“构造”的词语的数量及使用频率,从而反映高频词语的生成能力。由于语料量及词语数量庞大,所有的统计与分析都基于自动分词,不做人工干预。

表1:三家媒体共用总词数及三家媒体共用的高频词

年度	三家媒体共用总词种	总词语共用		三家媒体共用高频词	高频词语共用	
		词种数	比例(%)		词种数	比例(%)
2009年	193416	72641	37.56	7876	5696	72.32
2008年	189937		38.24	7621		74.74
2007年	176798		41.09	7740		73.59
2006年	143910		50.48	7575		75.19
2005年	106111		68.46	6624		85.99

4 操作流程

对常用词语构造能力的考察,我们分为两个部分,不仅要统计其构造词语的数量,同时也要考察其所构造的词语在语言使用中的地位与作用。

4.1 构词数量的统计

首先需要统计其所构造词语的数量。这里有一个现象值得注意,常用词语(即高频词表中的词语)本身就有整体与部分(这里的整体与部分,不是指语义层面的整体、部分,如:“手指”、“车轮”等;而是指词型上的整体与部分关系,如“人”和“人们”)的关系,所以我们采用词长“由大及小”的顺序来进行考查,即先考虑高频词语中词长最长的词的所构造词语的数量,而后再考虑更短词语的构造数量,在进行后者的考查时,不包括前者已有的构造数量。例如: S_1 、 S_2 都是高频词语,且 S_2 中包含了 S_1 (如: S_1 =“师”, S_2 =“老师”),则我们不再计算 S_1 在 S_2 中的出现次数。主要流程如下:

(1) 将高频词表 S_i 与总词表 S_j 进行比对,若 S_i 是 S_j 的真子集,则提取一条记录,最终形成一一对应的构造词表: S_i-S_j 。

(2) 找到 S_i 中的最长串集合 Set_n (表示词长为n的词语集合,以此类推),在本文高频词表中

* 对于常用词语的范围,人们是有不同看法的,我们认为可以以频率为基础,同时考虑词语在文本、时间轴、领域、地域等方面的散布度,从而划定其范围与等级。

词长最长为 7 字，只有一个元素：“中华人民共和国”。

(3) 判断 S_i 是否有词语包含“中华人民共和国”且不等同于“中华人民共和国”，若有，则从 S_i 提取出来，从而得到其所构造的词语。

(4) 在构造词表中，删除以下记录： S_i 中包含于“中华人民共和国”的词语，如“中华”、“人民”、“共和国”等，且其对应的构造词 S_i 与“中华人民共和国”对应的构造词相同，最终形成新的高频词表。

(5) 选定 Set_{n-1} ，在新的总词表中，重复第(3)、(4)条，直到 $n=1$ 。

最终得到的高频词语的构词数量表，其中包括高频词语 3938 条，也就是说，有 1758 条词语没有构成其他词语。

4.2 所构造词语的使用情况统计

由于受到分词软件的影响，命名实体的切分结果会有所不同，业界也认为有两种处理方法，一种是作为整体，如：“北京大学中文系/n”；一种是可以分开，“北京/n大学/n中文/n系/n”。这两种不同的处理方法各有利弊，专家学者们也是见仁见智，重点是取决于研究者的研究目标与使用需求。但这对于我们考查词语的构造能力有很大的影响（本文使用的分词软件多是采用前者的处理方法，从而得到一个切分单位）。因为语言资源监测语料库是来源于国内的各种主流媒体，对于组成社会语言生活的人物、地区、组织机构以及时间等都有大量、广泛地关注与报道，因此在频次较低的部分会涌现大量的表示人名、地名、组织机构名的词语，这些词语在词种数上占绝对优势，但在使用频率上却微乎其微^{*}。由此可见，仅仅依据高频词语的构造词语数量而断定其生成能力未免有些武断，我们需要进一步考查其所构造的词语在语言使用中的作用与地位。

因此本文还考查了这些高频词语所构词语的使用情况，将其所构词在五年语料库中的频率求和，得到所构词的使用情况，这样能更加客观地反映其构造能力。

5. 结果分析

词语只有在不断应用中才能存活并发展，它的生命力的一个突出表征是其构造其他词语的能力，我们认为不论是“部件词”、“类词缀”还是词缀，这些都是构造能力比较强的语言成分。我们可以在语言监测的基础上得到其构造能力统计结果，以期反映词汇变化发展的状况：按高频词所构造的词语数量和总频率分别排序，其前 20 位分列如下表 2：

由此可见，以上两种不同角度所得到词语的构造能力是不一样的，都有各自的特点。单纯考虑构造词语的数量，得到的构造能力较高的词语多为组织机构名（如：分、局）、人名（如：王、张、李）、地名（如：市、区）及时间词（如：年、月、日）以及其他（如：大、一、人）；而按所构造的词语总频率来衡量的话，更多是一些构成高频词的词缀或类词缀，这些词语有很强的生成作用，且其所构造的词语在语言使用中的作用和地位进一步突显。

在构词能力考查中，没有构造其他词语的高频词语也不少，有 1758 个。它们主要有：虚词、惯用语和双音合成词。虚词以语法意义或功能意义为主，其使用比较稳定，结构与意义都相对完

^{*} 从近几年的《中国语言生活报告》中字词语调查来看，字词的使用情况都在一定程度上验证了齐普夫定律，即少数高频字词会覆盖语料的大部分，在年度使用的词语中，高频词语占总词种数的比例不到 1%，低频词语词种数占绝对比例。

整。同时,合成词的结构以复合式、附加式*居多。从历史语言学的角度来看,合成词是顺应了汉语复音化的规律,如“穿越”古时多用“穿”或“越”,复音化后形成了联合型的双音词。以下列出其中使用频率较高的20条,括号内是其在总词表中的序号(按五年总使用频率降序排列)。痛苦(17)、高新技术(28)、配合(61)、同比(69)、上级(92)、资讯(181)、每月(219)、取出(260)、卡车(266)、报销(288)、严厉(350)、传说(397)、哪些(430)、是否(442)、两者(445)、加上(459)、穿越(557)、外出(689)、近年来(772)、被害人(784)、

表2: 高频词语的构造能力

词语	构词数量	词语	构词数量	词语	构词总频率	词语	构词总频率
大	897	日	579	国	2.636799	出	1.637032
人	819	水	575	人	2.396923	时	1.620675
年	809	中	556	有	2.053557	上	1.616238
一	802	山	552	一	1.992308	不	1.568948
不	781	小	530	中	1.945124	生	1.540724
月	766	点	520	大	1.875694	业	1.495712
市	688	分	518	年	1.723018	者	1.464911
子	618	区	477	行	1.652462	来	1.45849
性	590	者	473	为	1.648407	现	1.440981
天	587	王	455	发	1.639871	以	1.436913

综上,那些构成了其他词语的高频词,不仅仅有“部件词”,还有词缀、类词缀等。曾立英^[9]指出:类词缀和部件词有很多相似的地方,构词的能产性都很强,差别主要有两点:一是部件词意义比较实在,而类词缀意义比较虚化;二是类词缀有定位性,部件词没有定位性。

因此,本文进一步分析了类词缀的构词能力,采用了曾立英与尹海良的类词缀初始集,尹海良^[10]对比了北京大学历经10年研制的面向中文信息处理的《现代汉语语法信息词典》(1998)和教育部语用所《面向中文信息处理的词语切分与词性标注规范》对“前接成分”和“后接成分”的收录情况,结果显示二者的合集有72个;曾立英^[9]在定量考查与定性分析的基础上,总结了条类前缀23个,类后缀53个,共76个类词缀。上述两者的并集为104个(此处不区分同形异义或同形多音词,如词缀“家jia5”和“家jia55”都作为一个词缀进行统计)。

严格意义上讲,词缀有其界定的原则与方法,虽然语言学界莫衷一是,但也公认可以从语音的弱化、语义的虚化、位置的确定性、周遍的规则性^[11]等几个方面进行判定。另外,类词缀的判定则主要是强调其搭配的范围比较广泛,而对其意义上是否完全虚化并不深究。而本文所做的类词缀构词统计,并没有区分其位置的确定性,所以讨论的重点并没有放在这些高频词是否可以成为“词缀”,而只是重点分析其构造其他词语的能力。这些构造能力前20个的类词缀中,见表3。只有8个进入了前文表2(右侧)所列的按所构词总频率排序的前20个中:人、有、大、不、生、业、者、以。而在表2(右侧)靠前的词语中,表示数量、地域与时间的高频词较多,如“国、一、年、时”等,这与语料的特点也有关系,同时也应看到,构词的“部件”与词缀是有区别的。

此外,我们还应该看到,在104个类词缀中,只有94个类词缀构造了其他的词语,另有10个类词缀没构造其他词:乎、匠、炎、然、第、仪、半、零、艇、伪。

*这里的分类与术语引用黄伯荣、廖序东:《现代汉语》 高等教育出版社 增订三版 2002年

这归因于我们所采用的计算方法,当 S_1 是 S_2 的真子集时,由于只以较长的高频词语 S_2 作为构词单位计算,而 S_1 不再作为包括 S_2 的词语中的构词单位。例如:“然”的构词能力都被“突然、当然、然而、自然、天然气、偶然、天然、不然、必然、依然、既然”所替代。

表 3::类词缀的构词能力统计表

类词缀	构词数	类词缀	构词数	类词缀	总频率	类词缀	总频率
大	897	头	392	人	2.396923	方	1.206643
人	819	化	377	有	2.053557	场	1.174571
不	781	法	358	大	1.875694	工	1.032219
子	618	长	351	不	1.568948	过	1.026423
性	590	有	341	生	1.540724	机	0.966776
小	530	老	336	业	1.495712	面	0.945996
者	473	风	324	者	1.464911	力	0.93855
无	443	气	307	以	1.436913	可	0.93834
家	437	手	295	们	1.417535	子	0.929852
生	437	之	293	家	1.266538	体	0.915023

6 运用构词规律进行新词语提取的尝试

词缀、类词缀对于词语的提取有重要作用,我们可以尝试以词缀为锚点,辅之以其他的特征项,从而在大规模的语料库获得需要的词语。其中,在大规模语料中提取新词语是一项费时、费力的工程,已有的基于规则与基于统计的方法都不能保证能取得十分理想的结果。

语言监测中主要使用“机器+人工”的方法,即“全切分对比法”和“特征对比法”。因为当一个词语或一个意义刚产生的时候,人们使用它时往往会有一些形式上的特征,如用引号、括号等,而引号使用较为普遍^[12]。

因此,引号作为一个最重要的形式标记,可以较好地找到新词语的候选集,从 2006-2008 年的新词语提取结果可以看出,“特征对比法”有一定的效果,如下表 4。

表 4: 用“特征对比法”提取新词语效果一览表(平面媒体)

时间	监测结果	提取数量	相同	召回率
2008	108	751 805	44	0.4074
2007	214	168 237	163	0.7617
2006	135	137 492	102	0.7556

将提取的范围框定在平面媒体语料中,表中“监测结果”是指这三年发布的新词语在平面媒体中出现过的词语数量,“提取数量”是指用“特征对比法”提取后删去与往年相同的字串后的词语数,“相同”是指词语提取的结果与最后发布的新词语的比较。从中看出,召回率相对较高而精确率有待于提高。

而我们在 2009 年加入了类词缀的规则方法后,在一些新词族的提取上取了一定的效果。如:2009 年以“门”、“族”为结尾的新词语其类词缀的特征比较明显,用“全切分对比法”和“特征对比法”得到的初始集是“门”类有 1463 个,经人工挑选,不是新词语的有 887 个,主要是

以“入门”、“冷门”、“龙门”等结尾的词语；“族”类有1132个，216个不是新词语，主要是以“诛九族”、“家族”、“民族”结尾的词语。通过对新词语的结构特征与规律进行总结，可能进一步提高提取的准确率。新词语作为语言变化发展的一个重要外在体现，必然有一定的结构特点和规律。如对于“被+表主动的动词”、“楼+AA”等形式，可以在结构规律的基础上做提取。

7 总结与展望

基于2005-2009年五年语言资源监测的结果，本文对高频词语的构造能力做了一些统计与分析，这些基础的“构造词”既包括“部件词”，也包括类词缀、词缀等词语“构件”。通过对这些词语的组合、聚合规律的进一步探讨，可以帮助我们充分了解词库与词法的关系，了解和掌握更多的词法规则，“可以大大减少常用词库的冗余度，通过增加构词规则，并对‘部件词’增加属性描述：可以适用什么样的构词规则？这样的常用词库就不仅仅只包含数据库，也包含规则库，这也是词库的知识表示形式的发展。”^[13]此外，我们也应该看到：词语特征的分析与描写被放到非常重要的位置，走上了“大词库，小规则”之路^[14]。这些特征与结构规律的提炼，对于词汇知识的获取、词汇变化发展的监测，都有重要作用。

本文所做的工作只是简单的梳理与分析，“部件词”的获取、词法规则的总结、归纳与完善等工作都是长期而又艰巨的，本文所做的高频词构造能力的统计与分析，希望能够提供一定的参考，但如同“A里AB”“AABB”等重叠式紧缩为部件词“AB”，成语、习用语中“部件词”保留不变等原则的实施，还都需要人工与机器的相互合作。词语总是在规则范围内不断演变的，以常用词语为核心、辅之以灵活的规则与词法是否就可以以不变应万变？在统计方法及角度等方面，这还有待于我们进一步探索，本文仅是抛砖引玉，错谬之处，敬请各位专家批评指正。

参考文献：

- [1] 曾小兵,等.基于语言监测的数字化汉语教学的词汇更新.第七届中文电化教学国际研讨会论文集,2010.
- [2] 曾小兵,等.主流平面媒体中成语的使用情况及特征分析.语言教学与研究,待刊.
- [3] 赵元任.汉语口语语法.1968,见吕叔湘译本,北京:商务印书馆,2002版.
- [4] 朱德熙.语法讲义.北京:商务印书馆,1982.
- [5] 黄居仁,等.中文词汇网络:跨语言知识处理基础架构的设计理念与实践.中文信息学报,2010(2):14.
- [6] 赵小兵.基于动态流通语料库的现代汉语基本词汇自动识别与提取方法研究.北京语言大学博士论文,2007.
- [7] 黄居仁,张化瑞,俞士汶.基本词汇的预测与验证:由分布均匀度激发的研究构想.何大安,曾志朗编《永远的POLA——王士元先生七秩寿庆论文集》,台北:台湾中研院语言学研究所,2005.12.
- [8] 俞士汶,朱学锋,支流.基于计量研究的现代汉语常用词库的构建.见张普,王铁琨主编《中国语言资源论丛(一)》,北京:商务印书馆,2009.
- [9] 曾立英.现代汉语类词缀的定量与定性研究.世界汉语教学,2008(4):75-87
- [10] 尹海良.基于语料库的现代汉语词缀与派生词自动识别问题初探.语言文字应用,2010(1):125-134.
- [11] 董秀芳.汉语的词库与词法.北京:北京大学出版社,2005:34-41.
- [12] 国家语言资源监测与研究中心.《中国语言生活状况报告(2008)》下编.商务印书馆,2009:321.
- [13] 俞士汶,朱学锋,段慧明,李芸.《常用词表》考察与常用词库建设.待刊
- [14] 陆俭明.句法语义接口问题.外国语,2006(3):30-35